UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper
Economic Series 16-07
April 2016
ISSN 2340-5031

Departamento de Economía
Universidad Carlos III de Madrid
C/Madrid, 126 28903 Getafe (Spain)
Fax (34-91) 6249875

# Education Curriculum and Student Achievement: Theory and Evidence*

Vincenzo Andrietti† Xuejuan Su‡

**Abstract**

This paper proposes a theory of education curriculum and analyzes its distributional impact on student learning outcomes. Different curricula represent horizontal differentiation in the education technology, thus a curriculum change has distributional effects across students. We test the model using the quasi-natural experiment of the G8 reform in Germany. We find evidence of heterogeneous reform effects consistent with our theory. While the reform improves student test scores on average, such benefits are more pronounced for well-prepared students. In contrast, less-prepared students do not benefit from the reform.

***JEL Classification**:* I21, I28, D04

***Keywords**:* Education curriculum, horizontal differentiation, distributional effects, difference-in-differences, conditional quantile regression, unconditional quantile regression

**Acknowledgments:* We are grateful to Nicole Fortin for sharing the STATA codes with us. We thank seminar participants at the University of Alberta for their helpful comments. All remaining errors are our own.

† Università "G. d'Annunzio" di Chieti e Pescara, DiSFPEQ. E-mail: vincenzo.andrietti@unich.it
‡ University of Alberta, Department of Economics. E-mail: xuejuan1@ualberta.ca

# 1 Introduction

Students of different levels of preparation (or prior knowledge) have different learning needs. Given the hierarchical nature of a learning process, students need to comprehend, apply, and synthesize the basic materials before they can effectively learn more advanced ones. In other words, their human capital output from an earlier stage of learning becomes the input and determines the learning effectiveness at a subsequent stage. As a result, well-prepared students – i.e., those with good prior knowledge or, equivalently, high human capital output from previous learning – can learn new topics more quickly, while less-prepared students – i.e., those with poor prior knowledge – need remedial work on the old materials before they can embark on the new ones, so they learn the new topics more slowly. In this sense, when the pace of learning is ideally matched to a student's preparation, her learning effectiveness will improve, leading to better learning outcomes.

To capture this concept of an ideal match between the pace of learning and the preparation of a student, we propose a theoretical model of the education curriculum. More specifically, an education curriculum is characterized by two parameters: a progress rate, and a corresponding minimum threshold on student preparation. A more challenging curriculum is one with a fast progress rate and a high threshold, while a less challenging curriculum is one with a slow progress rate and a low threshold. Given different curriculum options, well-prepared students enjoy better learning outcomes under a more challenging curriculum, while the opposite is true for less-prepared students. Thus, different curricula represent "horizontal differentiation" in the education technology, and the "ideal" curriculum – namely, the one that maximizes a student's learning effectiveness – differs across students depending on their preparation levels.

The horizontal feature of the education curriculum stands in sharp contrast to other important factors of the education technology, which have been extensively studied in the literature. Many of those factors represent "vertical differentiation" in the education technology, so students share similar preferences when given different options. For example, school resources, class size, teacher quality (or teacher experience), and peer effects are all vertical features of the education technology. Since more resources, smaller clas-

ses, better teachers, and better peers all contribute to better learning outcomes, when given a choice, all students will prefer to have higher quality (along these dimensions) to lower quality, everything else the same. Unlike these vertical quality dimensions, not all students prefer to have a more (or less) challenging curriculum. Instead, they prefer different curricula depending on their preparation levels.

An immediate implication of the horizontal feature of the education curriculum is its distributional impact across students. When a number of students are subject to the same education curriculum, for example, because they live in the same state, attend the same school, or share the same class, the curriculum adopted in that state, school, or class applies to all of them despite their different preparation levels. Except for a lucky few where the match happens to be ideal, mismatches are likely to happen to many of the students. In particular, an over-match arises when the curriculum is too fast-paced (and accordingly imposes too high a threshold) for a student, and an under-match arises when the curriculum is too slow-paced for a student. Both types of mismatches are detrimental to the learning effectiveness of the students, and the more severe the mismatch, the more harm on their potential learning outcomes. The most extreme mismatches can be mitigated when a student repeats a grade (extreme over-match) or skips a grade (extreme under-match). However, moderate mismatches are likely to persist given that it is all but infeasible to individually customize the education curriculum to ideally suit each student's learning needs.

As a consequence, when there is a change in the education curriculum, there is similar distributional impact across students. For example, when the new curriculum is more challenging than the old one, it gets closer to the ideal curriculum for the well-prepared students, but farther away from that for the less-prepared students. As a result, well-prepared students benefit from the change and enjoy better learning outcomes, while the opposite is true for less-prepared students. In this sense, a curriculum change generates heterogeneous effects on student learning outcomes, and the relationship is monotonic in student preparation levels. More specifically, the relationship is monotonically increasing when the new curriculum is more challenging, and monotonically decreasing when the

new curriculum is less challenging.

We empirically test this model. More specifically, we are interested in finding out whether a curriculum change has heterogeneous effects on student learning outcomes, and whether the pattern of the heterogeneous effects is consistent with the theory. There are two major challenges in our empirical analysis. The first is that education curriculum, as characterized by the two parameters, is not directly measurable in the data. While most teachers intuitively adjust their pace of teaching to better serve their students (e.g., slowing down and doing a few extra practice questions when students seem to struggle with a topic), it is difficult to assign a numerical value to reflect such curriculum adjustment in practice. As a first step, we circumvent this problem by focusing on the ordinal rather than the cardinal comparison between two curricula: Namely, determining which curriculum is more challenging without deciding by how much. The second challenge is that education curriculum, as actually adopted by educators, is necessarily an endogenous choice. This choice naturally depends on both the distribution of student preparation and the objective function of the educator, both of which may be unobserved in the data. In this sense, cross-sectional variation of observed curricular differences can be confounded by important unobserved factors, making it unsuitable for identification purposes. We deal with this problem by relying on a quasi-natural experiment of the curriculum change.

For our empirical analysis, we take advantage of the G8 reform in Germany. This reform – implemented from 2001 to 2008 in most German states – compressed high school for the academic-track (*Gymnasium*) students from nine to eight years, while keeping the academic content required for graduation fixed. Namely, the reform requires the same amount of content being covered in a shorter time period, implying a faster progress rate and, accordingly, a higher preparation threshold. Thus, compared to the control (G9) states, the treated (G8) states have a more challenging curriculum. Furthermore, the G8 reform can be viewed as a quasi-natural experiment. It was implemented by states based mainly on considerations of the labor market conditions and demographic changes, with little focus on student learning outcomes directly. In this sense, the G8 reform can be viewed as an exogenous curriculum change, which allows us to identify the distributional

effects on student learning outcomes.

To measure student learning outcomes, we use five waves of PISA data containing their reading, mathematics, and science test scores at the end of the ninth grade. Since the pooled PISA data are repeated cross-sections rather than panel data, we have very limited information on student preparation when they entered high school. Two approaches are used in the empirical analysis. The first is the conventional difference-in-difference (DiD) approach, where time- and state-variation in the G8 reform implementation allow us to identify the average effect of the curriculum change. More importantly, we use some crude measure of student preparation to interact with the reform variable, and estimate the heterogeneous effects of the curriculum change as distinct average effects for two subgroups of students, the well-prepared and the less-prepared. The second is a quantile treatment effect approach, where we rely on distributional assumptions of the unobserved preparation variable. In particular, we use both the conventional quantile regression method (conditional quantile regression) and the recentered influence function method (unconditional quantile regression) in a nonlinear DiD setting. The results can be interpreted as the treatment effects at different quantiles of either the conditional distribution or the unconditional distribution of test scores respectively. The empirical evidence is broadly consistent with our theoretical predictions. While the G8 reform improves student test scores on average, the benefits are more pronounced for well-prepared students. In contrast, there is little evidence that less-prepared students benefit from the reform at all.

The rest of the paper is structured as follows. Section 2 reviews the related literature. Section 3 introduces the theoretical model of education curriculum and derives the model predictions. Section 4 presents the regression models that empirically test the theoretical predictions. Section 5 illustrates the natural experiment and the data exploited for the empirical analysis. Results are presented in Section 6. Section 7 offers concluding remarks.

## 2 Related literature

This paper is linked to several strands of the existing literature. On the theoretical side, there is a growing literature that focuses on the hierarchical nature of the education process, namely the human capital output from an earlier stage is an input for human capital accumulation and improves the learning effectiveness at a subsequent stage of education (see, for example, Ben-Porath, 1967; Lucas, 1988; Driskill and Horowitz, 2002; Su, 2004, 2006; Blankenau, 2005; Blankenau et al., 2007; Cunha and Heckman, 2007; Gilpin and Kaganovich, 2012). More specifically, a few of these studies (Su, 2004, 2006; Gilpin and Kaganovich, 2012; Kaganovich and Su, 2015) focus on the role of a curricular threshold as an important determinant in the education technology, and derive the implications of such a threshold on the aggregate efficiency and distributional equality.[1]

The paper is also related to the literature on how peer effects affect students' school choices and learning outcomes (see, among others, Rothschild and White, 1995; Winston, 1999; Epple and Romano, 1998, 2008; Epple et al., 2002, 2004, 2006). The peer effects literature captures a vertical feature of the education technology, namely the higher is the average quality of one's peers, the better off is a student in terms of her learning outcomes. Such a peer effect captures the "direct" externality that peers exert on a student's learning. In contrast, our paper focuses on the "indirect" peer effect: that is, one education curriculum is adopted to serve both the student and her peers. So even if the student does not directly benefit from having high-quality peers, he is nonetheless affected by the adopted education curriculum, which may be chosen to better serve her peers rather than himself. The two kinds of peer effects have drastically different implications. While the direct peer effect suggests that all students would prefer to have as high quality peers as possible, the indirect peer effect suggests that this is not necessarily optimal. For example, if a low ability (or less prepared) student were to attend a school with predominantly high ability (or well prepared) students, he would find its curriculum (which is geared toward the high ability students) overly challenging and hence experience a negative impact on

[1]For models of academic standards as a requirement on the education outcome, see Costrell (1994, 1997); Betts (2008); Eisenkopf and Wohlschlegel (2012).

her learning outcomes.

On the empirical side, there is a large existing literature estimating the impact of various vertical measures of the education technology, such as school quality and school resources (Card and Krueger, 1992; Currie and Dee, 1995, 2000; Hanushek, 1997, 2006; Jacob and Lefgren, 2004), class size (Angrist and Lavy, 1999; Hoxby, 2000; Krueger, 2003; Ding and Lehrer, 2010; Chetty et al., 2011), teacher quality (Angrist and Lavy, 2001; Rivkin et al., 2005; Clotfelter et al., 2006; Aaronson et al., 2007; Rothstein, 2010; Carrell and West, 2010; Mueller, 2013), and peer effects (Evans et al., 1992; Sacerdote, 2001; Zimmerman, 2003; Angrist and Lang, 2004; Arcidiacono and Nicholson, 2005; Lyle, 2007; Carrell et al., 2008, 2009). This literature tends to focus on the average treatment effect associated with the change in one of these vertical measures, since economic theory provides an unambiguous prediction as to the qualitative impact (the direction) of such a change on student learning outcomes. On the other hand, the focus is typically not on the distributional effect, since economic theory tends to be ambivalent as to how the quantitative impact of such a vertical change would vary across students.

There is also a small but fast growing empirical literature that focuses on the distributional effect of matches between students and schools. For example, Light and Strayer (2000) examine whether the match between student ability and college quality affects the student's college graduation rate. They find that while high-ability students are on average more likely to graduate from college than low-ability students, as expected, students of all ability levels are more likely to graduate if they attend colleges with quality level matching their ability level. In other words, high-ability students are more likely to graduate when attending high-quality rather than low-quality colleges, while the opposite is true for low-ability students. More recently, Arcidiacono et al. (2011) examine whether affirmative action leads to mismatch between lower-ability students and highly selective schools. They find evidence that, compared to the school, students are worse at predicting their post-enrollment achievement based on initial preparation. Thus, affirmative action can result in mismatches: had students known that they would perform worse than expected, they could have chosen a different (less selective) school. Similarly, Arcidiacono

et al. (2016) examine the difference in the graduation rates for minority science students across University of California campuses under affirmative action policies. They find that less-prepared minority students at higher-ranked campuses had lower persistence rates in science and took longer to graduate. Again, affirmative action can result in mismatches: had these minority students attended lower-ranked campuses and hence had they been better matched to universities according to their initial preparation, they would have reached higher graduation rates in STEM fields. This line of evidence – that lower-ability students enjoy better learning outcomes in less selective schools – hints at the existence of important factors that "horizontally" differentiate less selective schools from more selective schools. Our paper provides a theoretical explanation of the education curriculum as one possible horizontal factor.

The paper that is most closely related to our paper is Duflo et al. (2011). In this study, they examine whether academic tracking helps or hurts low-ability students. Using randomized experimental data from Kenya, they find that tracking students by prior-achievement raises scores for all students, even those assigned to lower achieving peers. To interpret these results, they argue that tracking allows teachers to better tailor their instruction level, and lower-achieving pupils are particularly likely to benefit from tracking when teachers have incentives to teach to the top of the distribution. Similar to our paper, their model allows both a "direct" peer effect (student-to-student spillovers) and an "indirect" peer effect, where the indirect effect arises when the composition of the class affects teacher effort as well as the target level of teacher instruction. Unlike our paper, their model does not allow the trade-off between the target level and the pace of learning, the two related parameters of the education curriculum in our model. Instead, they model the pace of learning as a result of teacher effort, which can be changed independently of the target level of instruction. In a sense, our paper can be viewed as moving along a given efficiency frontier of the education technology consisting of different curricula, while their paper can be viewed as improving the efficiency frontier when changes in the teaching environment (tracking) and stronger incentives (contract teachers) induce higher levels of teacher effort, which again is a vertical measure of the education technology. Alternatively

speaking, in our model, high-ability students have an absolute advantage over low-ability students in their learning effectiveness (value-added human capital) regardless of the curriculum, even though their comparative advantage is in more challenging curricula. In their model, high-ability students have no absolute advantage over low-ability students *per se*, and their learning effectiveness will be the same as long as the target level of teacher instruction is at the same distance away from their initial preparation.

# 3 A model of education curriculum

Consider an economy with heterogeneous students. Students differ by their initial preparation $q_i \in [\underline{q}, \overline{q}]$. We will discuss the distribution of student preparation later. Depending on whether the focus of the analysis is at the micro level (class or school) or macro level (state or country), it is more convenient to treat the student distribution alternatively as discrete or continuous, but the main results remain robust regardless of the particular distribution under consideration.

## 3.1 Education curriculum

An *education curriculum* is defined by two parameters: A progress rate $A$ which captures the pace of learning, and a corresponding curricular standard $c(A)$ that puts the minimum requirement on student initial preparation. Thus, when a student with initial preparation $q$ (we omit the subscript of $q_i$ when there is no risk of confusion) studies under the curriculum $(A, c(A))$, her human capital output $h$ from a period of study is

$$h = \begin{cases} (1-\lambda)q & \text{if} \quad q \leq c(A), \\ (1-\lambda)q \quad +A(q-c(A)) & \text{if} \quad q > c(A). \end{cases} \tag{1}$$

Namely, a student's preparation (or existing human capital) $q$ depreciates at the rate $\lambda \geq 0$, so only the undepreciated part $(1-\lambda)q$ is kept after the study period. Furthermore, when the preparation level fails to meet the threshold $c(A)$, the student does not benefit from the learning process and accumulates zero from the study period. On the other

hand, when the preparation level surpasses the threshold, the value-added human capital from the learning process is $A(q - c(A))$.

It is immediately clear that if there was a curriculum $(A, c(A))$ with a very large value for $A$ and a very small value for $c(A)$, it would give large benefit to students of almost any level of preparation. In a world of trade-offs, such a technology is unlikely to be feasible. At the efficiency frontier, curriculum choices involve a tradeoff. That is, larger values for $A$ (faster pace of learning) requires larger values for $c(A)$ (higher requirement on initial preparation).

**Assumption 1** *Let the curricular threshold $c(A)$ be a differentiable function with $c'(A) > 0$.*

We maintain Assumption 1 hereinafter. A direct implication of the specification of the education curriculum is that well-prepared students have an absolute advantage over less-prepared students in a given curriculum, a distinguishing feature of our model from Duflo et al. (2011).

**Proposition 1** *For any given curriculum, well-prepared students have an absolute advantage over less-prepared students. Namely, when $q' > q > c(A)$, $A(q' - c(A)) > A(q - c(A))$.*

The proof follows directly from (1). Note that well-prepared students not only enjoy an absolute advantage over less-prepared students, they also enjoy increasing marginal returns to their preparation: Comparing two students with preparation $q' > q > c(A)$, we have not only $A(q' - c(A)) > A(q - c(A)) > 0$, but also $\frac{A(q'-c(A))}{A(q-c(A))} > \frac{q'}{q} > 1$.

## 3.2 Ideal curriculum for a student

If an educator were able to customize the education curriculum to serve the individual learning needs of a given student with preparation $q$, the educator would have chosen a curriculum that maximizes the student's human capital output $h$ according to (1). This optimal choice would be the ideal curriculum for this given student. For example, if we make the assumption that $c = CA^r$ with $r > 0$, the ideal curriculum for a student with

preparation $q$ is $A^*(q) = \text{argmax}\, A(q - CA^r) = (\frac{q}{C(r+1)})^{1/r}$, which is strictly increasing in $q$. More generally, without a specific functional form for $c(A)$, we may not explicitly solve for the ideal curriculum $A^*(q)$. Nonetheless, it is implicitly defined as the solution to the first-order equation:

$$q - c(A) - Ac'(A) = 0 \tag{2}$$

assuming that the second-order sufficient condition is also satisfied, namely $-2c'(A) - Ac''(A) < 0$. Applying the Implicit Function Theorem to (2), and then invoking the second-order sufficient condition, we have the following result:

**Proposition 2** *The ideal curriculum for a student is strictly increasing in her initial preparation $q$.*

In other words, regardless of the particular function form that links the threshold $c(A)$ to the progress rate $A$, well-prepared students always benefit more and hence enjoy a comparative advantage in faster-paced (more challenging) curricula, while less-prepared students always benefit more and hence enjoy a comparative advantage in slower-paced (less challenging) curricula. This is the core feature of the "horizontal" aspect of the education curriculum, that different students would prepare to have different curricula to best suite their learning needs.

## 3.3 Implemented curriculum

In practice, it is typically infeasible for an educator to customize the education curriculum to serve the individual learning needs of a given student. Instead, a number of students may enroll in the same school or attend the same class, and hence be exposed to a common education curriculum despite their different preparation levels. When this is the case, the implemented curriculum may not be ideal for all but a few students. Instead, it can be too fast-paced for some students, and too slow-paced for others.

In this paper, we do not explicitly model how a curriculum gets chosen. In principle, the optimal curriculum choice can be derived as the optimal solution from maximizing the objective function of a teacher, a school, or a society. For example, consider a collection

of $N$ students with different preparation levels $q_i$, where $q_1 \leq q_2 \leq ... \leq q_N$. Assuming that the objective function is linear in each student's human capital outcome from the chosen curriculum, the optimal curriculum can be expressed as

$$A^*_{com} \;=\; \text{argmax}\, \Sigma^N_{i=1}\gamma_i h_i \quad s.t.(1), \quad \Sigma^N_{i=1}\gamma_i = 1, \tag{3}$$

where $\gamma_i$ is the relative weight the educator assigns to student $i$. So, similar to the interpretation of a social welfare function, when $\gamma_i = \gamma$ for all $i$, the educator is "utilitarian" and treats all students with equal concern; when $\gamma_1 \geq \gamma_2 \geq ... \geq \gamma_N$, the educator is more concerned about the less-prepared students ("no child left behind"); and when $\gamma_1 \leq \gamma_2 \leq ... \leq \gamma_N$, the educator is more concerned about the more-prepared students.

As is obvious from the set up, the optimal curriculum chosen by the educator for this collection of students depends critically on two factors: the distribution of the student preparation, and the relative weights assigned to different students. Here we can only characterize the comparative statics of the optimal curriculum in a few special cases. For instance, holding the set of relative weights fixed, the optimal curriculum becomes more challenging when the distribution of student preparation shifts to the right, i.e., when the new distribution is first-order stochastically dominant over the old distribution. Similarly, holding the distribution of student preparation fixed, the optimal curriculum becomes more challenging when there is a shift of the relative weights from less-prepared students to more-prepared students, i.e., from $\gamma_i$ to $\gamma_j$ when $i < j$. On the other hand, when there are more complex changes in either the distribution of student preparation or the set of relative weights, the optimal curriculum will depend critically on the quantitative comparison of the changes, and we cannot qualitatively characterize the comparative statics. An immediate implication is that, to empirically identify and estimate the impact of the education curriculum on student achievement, cross-sectional variation of the education curriculum in an observational dataset is of limited value. Unless we have perfection information on the distribution of student preparation as well as the implemented curriculum, self-selection bias poses a major challenge. As will become clear in our empirical section, we overcome this hurdle by relying on a quasi-natural experiment arising from a poli-

cy change in the education curriculum, where we argue that the distribution of student preparation remains stable before and after the curriculum change.

At the same time, it is reasonable to assume that regardless of the specific objective function, an implemented curriculum has to fall within the two extreme curricula: the ideal curriculum for the least-prepared student ($q_i = \underline{q}$) and that for the most-prepared student ($q_i = \overline{q}$). Otherwise, any curriculum with $A < A^*(\underline{q})$ is strictly Pareto dominated by $A^*(\underline{q})$, and any curriculum with $A > A^*(\overline{q})$ is strictly Pareto dominated by $A^*(\overline{q})$. So an implemented curriculum has to fall in the interval $A \in [A^*(\underline{q}), A^*(\overline{q})]$. Comparing two different curricula in this interval, we have the following stratification result:

**Proposition 3** *Consider two curricula $(A, c)$ and $(A', c')$ with $A^*(\underline{q}) < A < A' < A^*(\overline{q})$. There exists a cutoff level $\hat{q} \equiv \frac{A'c' - Ac}{A' - A} \in (\underline{q}, \overline{q})$ such that students with $q_i = \hat{q}$ accumulate the same level of human capital under the two curricula; students with $\underline{q} < q_i < \hat{q}$ accumulate higher levels of human capital under the old curriculum $(A, c)$ than the new; and students with $\hat{q} < q_i < \overline{q}$ accumulate higher levels of human capital under the new curriculum $(A', c')$ than the old.*

## 3.4 Grade repetition

The basic curriculum model can be easily extended to understand the role of grade repetition. More specifically, each grade has its own curriculum, namely the grade-specific progress rate $A_g$ and the grade-specific threshold level $c(A_g)$, where the subscript $g$ indicates the particular grade under consideration. When students finish grade $g$ and move up to grade $g + 1$, their human capital output from grade $g$, namely $h^g$, becomes the input for their learning process at grade $g + 1$, namely $q^{g+1}$. In other words, the entire learning process can be modeled as a series of hierarchical curricula:

$$h^{g+1} = \begin{cases} (1-\lambda)h^g & \text{if } h^g \leq c(A_{g+1}), \\ (1-\lambda)h^g \quad +A_{g+1}(h^g - c(A_{g+1})) & \text{if } h^g > c(A_{g+1}). \end{cases} \tag{4}$$

with the initial level $h^0$ being the student innate ability.

Now consider the situation of a student who, due to various reasons, finishes grade $g$ with a low level of $h^g$. The educator (or the student himself) faces the following decision: which is more beneficial to the student, moving on to the next grade or repeating the current grade. From (4), it is straight-forward to see that if $h^g \leq c(A_{g+1})$, the student will not be able to benefit from the next grade, so her only option is to repeat the current grade $g$. On the other hand, even if $c(A_{g+1}) < h^g < \frac{A_{g+1}c(A_{g+1}) - A_g c(A_g)}{A_{g+1} - A_g}$, the student still benefits more from repeating the current grade than moving on to the next grade, i.e., accumulating more human capital from attending grade $g$ instead of $g+1$. For this reason, as will become clear in the next Section, we use grade repetition as a crude indicator for students who are less-prepared.

# 4 Regression models

Of course, many other factors beyond the education curriculum – e.g., class size, teacher quality, parental engagement, tutoring, just to name a few – affect a student human capital outcome within a learning period. Our main focus in this paper is on the role of the curriculum, so all the other factors are used as control variables to the extent we have data. In particular, we are interested in a regression model as follows:

$$\begin{aligned} h_{ist} &= \alpha q_{ist} + A_{st}(q_{ist} - c(A_{st})) + \delta X_{ist} + \gamma_s + \eta_t + \epsilon_{ist} \\ &= \alpha q_{ist} + q_{ist} * A_{st} - f(A)_{st} + \delta X_{ist} + \gamma_s + \eta_t + \epsilon_{ist}, \end{aligned} \tag{5}$$

where $f(A)_{st} = A_{st} * c(A_{st})$ is a general function of the progress rate $A_{st}$. The dependent variable $h_{ist}$ is the test score for student $i$ in state $s$ and year $t$. On the right hand side, the first term represents the undepreciated human capital level. The next term is the value-added human capital output from the learning process under the given curriculum $A_{st}$ and $c(A)_{st}$, which can be expressed as an interaction term between the curriculum and the student preparation, and a term involving only the curriculum. The vector $X_{ist}$ is a set of other control variables that may affect learning outcomes, including student characteristics, family background, and school characteristics. We also allow for state

fixed effects $\gamma_s$ and year fixed effects $\eta_t$. Finally, $\epsilon_{ist}$ is the error term.

Note that our focal interest is on the two variables $q_{ist}$ and $A_{st}$, neither of which is directly measured. As will be discussed in the next section, the quasi-natural experiment of the G8 reform directly translates into an increase in $A_{st}$. Even if we do not have a quantitative measurement of the increase in the curriculum threshold, we can argue qualitatively that the G8 reform corresponds to an increase in the curriculum from one level (for all states under the old G9 regime) to another, higher, level (under the new G8 regime). In this sense, even though in theory the curriculum can take any value in the relevant range, for our empirical analysis we are only considering the difference between two particular curricula implemented in the G9 and G8 regimes. From here onward, we denote the curriculum associated with the old G9 regime as $A_o$ and $c_o$, and the curriculum under the G8 reform as $A_n$ and $c_n$, where $A_o < A_n$ and $c_o < c_n$. On the other hand, when there is only limited information on student preparation $q_i$, we use two different econometric approaches as described in detail below.

## 4.1 Difference-in-differences

Suppose student initial preparations are not perfectly observed. As long as the distribution of $q_i$ remains stable before and after the curriculum change, we can treat it as an unobserved variable and integrate it out to estimate an average impact of the curriculum change. More specifically, consider two states $s$ and $r$ in two years $t$ and $w$, such that $A_{st} = A_{rt} = A_{rw} = A_o$ while $A_{sw} = A_n$, namely state $s$ started with the original G9 regime in year $t$ but implemented the G8 reform in year $w$, while state $r$ maintained the G9 regime in both years. Inserting these particular values of the applicable curricula into (5) and then taking the difference-in-differences, we have

$$(h_{isw} - h_{jst}) - (h_{krw} - h_{lrt}) = \alpha((q_i - q_j) - (q_k - q_l)) + ((A_{sw}q_i - A_{st}q_j) - (A_{rw}q_k - A_{rt}q_l))$$
$$- (f(A_n) - f(A_o)) + \delta((X_{isw} - X_{jst}) - (X_{krw} - X_{lrt})) + (\epsilon_{isw} - \epsilon_{jst}) - (\epsilon_{krw} - \epsilon_{lrt}) \quad (6)$$

Even if we do not observe individual $q$ directly, as long as the distribution of student preparation in a given state stays the same over time, we can integrate them out to get the expected value or the average level for the student population.

**Assumption 2** *Let $\phi_{st}(\cdot)$ be the distribution density function of student initial preparations for state $s$ in year $t$, such that $\phi_{st}(\cdot) = \phi_s(\cdot)$ with mean $\mu_s$.*

Then, conditional on the observed variables $\mathbf{X}$, the average reform effect can be expressed as follows:

$$DiD(h) = \mu_s(A_n - A_o) - (f(A_n) - f(A_o)). \tag{7}$$

Namely, after controlling for the impact of the observed variables on the test score, our DiD approach allows us to estimate the average reform impact on student test scores even when we do not observe student initial preparations. The validity of the DiD approach relies on the assumption of a stable distribution for a given state over the years.

Similarly, we can break the overall average effect in (7) into two average effects, depending on whether a student's preparation is above or below a given cutoff level, a crude binary measure of student preparation. If the cutoff level happens to be $\hat{q}$ (as defined in Proposition 3), we have:

$$DiD(h^+) = \int_{q>\hat{q}} \phi_s(q)dq \times (A_n - A_o) - (f(A_n) - f(A_o)) > 0,$$

and

$$DiD(h^-) = \int_{q\leq\hat{q}} \phi_s(q)dq \times (A_n - A_o) - (f(A_n) - f(A_o)) < 0.$$

Alternatively, for any given cutoff level $\overline{q}$, even though the signs of $DiD(h^+)$ and $DiD(h^-)$ are not guaranteed to be positive and negative, we still expect the relationship $DiD(h^+) > DiD(h^-)$, or equivalently, $DiD(h^+) - DiD(h^-) > 0$.

Based on this, our DiD model is estimated by the following equations. For the overall average effect, we have:

$$h_{ist} = \beta_0 + \beta_1 G8_{st} + \delta X_{ist} + \gamma_s + \eta_t + \epsilon_{ist}, \tag{8}$$

where $h_{ist}$ is the (standardized) PISA reading, mathematics, or science score measured in year $t$ for an academic-track student $i$ in state $s$. $G8_{st}$ is the G8 reform indicator which equals one if a student observed in year $t$ and in state $s$ belongs to the cohort treated by the G8 reform in that state, and zero otherwise.

For the average effects within two subgroups of students, we have:

$$h_{ist} = \beta_0 + \beta_2 G8_{st} \times I(q_{ist} > \hat{q}) + \beta_3 G8_{st} \times I(q_{ist} \leq \hat{q}) + \delta X_{ist} + \gamma_s + \eta_t + \epsilon_{ist}. \tag{9}$$

Our main interest is the relationship between $\beta_2$ and $\beta_3$. Our theory predicts that $\beta_2 > \beta_3$, which will be tested in the data. In contrast, $\beta_1$ is a weighted average of $\beta_2$ and $\beta_3$, and is equivalent to the average reform effect across all students.

## 4.2 Quantile Regressions

An alternative approach to deal with the unobserved $q$ problem is quantile regression, which allows us to examine potentially heterogeneous effects the reform has at different locations of the outcome distribution. More specifically, the conventional quantile regression approach relies on the common distribution assumption, that is, not only should the distribution of the unobserved variable $q$ remains stable overtime, the distribution has to be the same across the treated and the control states. When this is the case, students at the $\tau$-th quantile would have exactly the same preparation $q_\tau$, regardless of whether they are in a treated or control state, before or after the treatment. Thus, holding all other observed variables constant, test score difference at a given quantile $\tau$ between the treated and the control state before and after the treatment can be attributed to the treatment itself, namely the G8 reform effect. We call this the quantile difference-in-difference

method (QDiD). To implement QDiD, we estimate the following quantile model:

$$h_{\tau,st} = \beta_{\tau,0} + \beta_{\tau,1} G8_{st} + \delta_\tau X_{ist} + \gamma_{\tau,s} + \eta_{\tau,t}, \tag{10}$$

where $h_{\tau,st}$ is the test score at a given quantile $\tau$ in state $s$ and year $t$. Note also that all the parameters are quantile-specific. In particular, the quantile-specific $\beta_{\tau,1}$ represents the treatment effect of the G8 reform at the particular quantile $\tau$.

However, the common distribution assumption is a well-known limitation of the quantile regression approach, which cannot be expected to hold in general. Without this assumption, the distribution of $q$ can be different in a treated state from that in a control state, and the test score difference at a given quantile may be attributed to either the preparation difference or the G8 reform itself, making the control state not a valid counterfactual for the treated state. To address this concern, we also use the Recentered Influence Function (RIF) method recently developed by Firpo et al. (2009), which explicitly relaxes the common distribution assumption.[2] More specifically, when the observed outcomes (in this case, test scores) vary monotonically with the unobserved variable (in this case, student preparation), RIF for the $\tau$th quantile as:

$$RIF(Y; q_\tau) = q_\tau + \frac{\tau - \mathbb{1}\{Y \leq q_\tau\}}{f_Y(q_\tau)}, \tag{11}$$

where $\mathbb{1}\{Y \leq \tau\}$ is an indicator variable that takes the value of 1 if $Y \leq q_\tau$ and 0 otherwise, and $f_Y(q_\tau)$ is the marginal distribution of $Y$ around the value of $q_\tau$. It has been shown that a RIF regression – defined for the $\tau$-th quantile as $E[RIF(Y; q_\tau)|X] = m_\tau(X) \approx X'\beta_\tau$ – leads to a consistent estimate of the unconditional quantile treatment effect.[3]

For our analysis, instead of examining students at the same quantile across states and years (as in the QDiD case), the RIF method compares students with the same test score and hence located at potentially different quantiles of the distributions across states

[2]Given its flexibility, the RIF method has recently been applied to analyze a range of issues such as cigarette taxes (Maclean et al., 2014) and child care (Havnes and Mogstad, 2015).

[3]See Firpo et al. (2009); Borah and Basu (2013).

and years. More specifically, consider a year before the reform, let us call it year $t$. For a given test score $h$, we can determine the corresponding quantiles $\tau_{st}$ in state $s$ and $\tau_{rt}$ in state $r$. Next, moving on to a subsequent year $w$ where state $s$ has implemented the reform but not state $r$. Again, for the same test score $h$, we can determine the corresponding quantiles $\tau_{sw}$ and $\tau_{rw}$ respectively. The impact of the G8 reform, measured as the change in the population shares that remain below the given test score $h$, is then given by $-((\tau_{sw} - \tau_{st}) - (\tau_{rw} - \tau_{rt}))$. This probability difference is then divided by a kernel estimate of the joint density of test scores at the level $h$ to arrive at the associated treatment effect. We call this the RIF-DiD method.[4] For the RIF-DiD method to work, the distribution of the unobserved variable (student preparation) can be different across treated and control states, as long as it remains stable over time within each state. This is much less restrictive compared to the common distribution assumption required for the QDiD method.

Besides the difference in the underlying distribution, there is also a difference in terms of interpretation of the estimation results. More specifically, the QDiD estimates can be viewed as the conditional quantile treatment effect, where heterogeneity in the observed variables implies potentially many different distributions. This matches closely with our theoretical model interpretation of the treatment effect due to a curriculum change, holding everything else constant. However, the conditional quantile treatment effect can be quite sensitive to the variables that it conditions on (Borah and Basu, 2013; Maclean et al., 2014). On the other hand, the RIF-DiD estimates can be viewed as the unconditional quantile treatment effect, where the many different conditional distributions are aggregated into one common unconditional distribution, given the realized values of the observed variables in the data. As a result, the unconditional quantile treatment effect is easily interpreted as that applicable to the entire student distribution. However, its link to our theoretical prediction of the curriculum effect is less direct. For example, suppose in the data, well-prepared students in treated states concentrate more heavily in middle

[4]We implement the RIF-DiD estimation procedure using the STATA ado file `rifreg` – downloaded from http://faculty.arts.ubc.ca/nfortin/datahead.html (last accessed December, 2015). The RIF is computed using a Gaussian kernel with an optimal bandwidth.

quantiles instead of top quantiles of the unconditional test score distribution, possibly due to individual heterogeneity in observed variables such as family background. In this case, the QDiD method can still accurately estimate the effect of a curriculum change, controlling for the differences in observed variables. On the other hand, the RIF-DiD method will conclude that the reform has a stronger effect in middle quantiles than top quantiles, because it reflects both within-group difference (that is, groups of the same family background) and between-group difference (groups of different family background).

A further limitation of both the QDiD and the RIF-DiD method is that, despite the importance of clustering standard errors at the treatment (state) level to avoid overstating precision (Bertrand et al., 2004) is widely recognized, a statistically valid method to cluster standard errors has not been developed yet. This is further complicated by the sampling weights associated with the observations in the complex survey design. As a result, we can only report the standard error for QDiD assuming i.i.d. residues, while that for RIF-DiD is bootstrapped using 200 repetitions.

# 5 Data

## 5.1 The policy variable - G8 reform

Educational policy in the Federal Republic of Germany is under the responsibility of the sixteen federal states. Children typically enroll in primary school at the age of six, and continue on to secondary school after four years. At the beginning of grade 5, students are tracked into three types of school: The basic-track school (*Hauptschule*) and the middle-track school (*Realschule*) provide vocational oriented schooling through grade 9 or 10; the academic-track high school (*Gymnasium*) leads to university entrance qualification called "*Abitur*".

Beginning in 2001, most German states introduced the so called G8 reform. The length of the academic-track curriculum was shortened by one year (from 9 to 8), but the total amount of curricular content to be covered as a graduation requirement was held fixed. As a consequence, the G8 curriculum has a faster progress rate – and, implicitly,

a higher curricular threshold – than the G9 curriculum. Using the terminology of our theoretical model, the G8 regime adopted a more challenging curriculum. Figure 1 offers a visual summary of the G8 reform implementation over time and across states. We refer to Andrietti (2015, 2016) for a detailed discussion of the G8 reform implementation and for a definition of the G8 policy variable.

For our analysis, the G8 reform can be viewed as a quasi-natural experiment: namely it was mostly driven by considerations of the labor market conditions and demographic changes. For example, in earlier policy discussions, then-federal secretary of education Jürgen Mölleman called on stakeholders to engage in deliberations on the subject "Twelve years (including primary school) to the *Abitur*." In his opinion, Gymnasium grades should be reduced from 9 to 8 years for the following reasons: "[German] graduates are two to three years older than their peers against whom they compete for jobs in the European labor market. ... German pension systems and demographics (characterized by a significant fraction of senior, retired citizens) cannot support such a late start of employment by young adults. ... Students reach the age of majority at 18 and should have completed post-secondary schooling by then, especially since the motivation for studying decreases with age. At age 25, they should have completed college, including military or civil service, and should have reached full social and economic independence. ... In addition, many schools do not fully utilize the 13th school year. Therefore, a decrease in education quality [associated with reform] can be avoided through more intensive instruction in smaller classes and, possibly, all-day instruction programs. (Translation by author)" (Wiater, 1996). Similarly, when the reform was actually implemented, its was implemented for similar reasons: "Following a change of government, Saarland was the first West German state to reduce the number of grades taken to reach *Abitur* from 13 to 12, effective academic year 2001/02. Driving this change was the supposed disadvantage of Saarland's graduates when entering the labor market caused by Germany's comparatively long schooling duration. ... As mentioned earlier, reducing the number of years of education is one of several measures aimed at lowering the age at which academically qualified workers enter the labor force, which is regarded as too high when

compared internationally and, in light of the rising demand for highly educated workers in a globalizing world, is expected to result in a competitive disadvantage for German university graduates, and hence for Germany itself. ... In order to protect social insurance systems, the palpable aging of the population, coupled with the simultaneous decline in births and population, necessitates an earlier entry of young adults into a longer phase of gainful employment. (Translation by author)" See Kühn et al. (2013) for more detailed discussions.

From a student perspective, the curriculum change is exogenous, and DiD is a suitable method to estimate the average treatment effects among subgroups of students. In a sense, the G8 reform is a policy-induced instrument for the actual curricula implemented at different schools across different years. Furthermore, since the G8 reform happened at the state level, the stability of their underlying preparation distribution is more likely to hold than that at more disaggregate levels, as there is very limited student mobility across state borders. Finally, Andrietti (2015) provides further support to the quasi-experimental nature of the G8 reform, documenting that high school enrollment patterns did not change in response to the introduction of its more challenging curriculum.

## 5.2 PISA data

The empirical analysis is based on a dataset that pools the first five waves of PISA assessment (2000, 2003, 2006, 2009, and 2012) for Germany.[5] While PISA is conducted by the OECD in a number of countries sampling 15-year-old students, independent of grade, national grade- and/or age-based extensions of the study were conducted in Germany for all PISA cycles, with the purpose of providing a sample large enough to allow comparisons between the different federal states. Given that the age-based PISA 2009 sample has not been released with state identifiers, our empirical analysis is based on grade-9 samples. In particular, our samples include all ninth-graders enrolled in academic-track high schools, with a valid test score assessment and with non-missing values on grade repetition.[6]

[5] Baumert et al. (2009); Prenzel et al. (2007, 2010); Klieme et al. (2013); Prenzel et al. (2015)

[6] Rather than dropping a small number of observations where information is missing on other background variables, we recode missing values to zero, and define missing values indicators for the variables included in a specification. In results available upon request, we find, however, that our main results are

It is worth pointing out that a sample of ninth-graders, like the one we use, includes high school grade repeaters. The latter spend one extra year in high school, compared to everyone else. The grade repetition essentially slows down the progress rate for these students, i.e., they learn the same amount of academic content with extra time and more slowly than dictated by the G8 reform. The extra year of schooling or remedial work adds up to what they would otherwise have achieved in the same amount of time as everyone else. Thus, grade repetition leads to potential upward bias in the reform effect for the grade repeaters. Despite the potential upward bias, if we still find evidence that grade repeaters benefit less from the G8 reform compared to the non-repeaters, we will know that the performance gap between the two subgroups of students will be even bigger when the upward bias is properly accounted for. In other words, our estimate offers a lower bound of the true effect.

PISA tests cover three different subjects (reading, mathematics, and science), assessing a range of relevant skills and competencies. Each subject is tested using a broad sample of tasks with differing levels of difficulty to represent a coherent and comprehensive indicator of the continuum of students' abilities.[7] An issue related to the pooled nature of our data is the comparability of subject-specific student assessments across PISA cycles. While reading assessments are comparable across all cycles, mathematics and science assessments underwent major revisions in 2003 and 2006 respectively, the first time they were considered the main subject. As a robustness check, we use both the full sample (all five waves) and the truncated sample (excluding 2000 for mathematics, excluding 2000 and 2003 for science) for estimation.

## 5.3 Control variables

Two groups of variables, defined at the student- and school-level, are employed as controls in the empirical analysis. Descriptive statistics on these variables are reported in Table

robust to the exclusions of missing values observations.

[7]Using item response theory, PISA maps student performance in each subject on a scale with an international mean of 500 and a standard deviation of 100 across the OECD countries included in the study. The scores are averages of plausible values, which are drawn from a distribution of values that a student with the given amount of correct answers could achieve as a test score (OECD, 2012).

1.

Student controls include a set of demographic and socio-economic characteristics, as well as a grade retention dummy that controls for different schooling experiences. The demographic characteristics include a dummy indicating female students and a quadratic age term (in months) that controls for potential age/maturation effects. The socio-economic characteristics include an indicator for the number of books at home, two indicators for parents' highest educational level (ISCED), as well as the Highest International Socio-Economic Index (HISEI), which uses the higher of the two parents' ISEI scores or the only available parent's ISEI score. There are also variables indicating a student's migration background, namely whether the student was born in a foreign country, whether a foreign language is spoken at home, and whether at least one of the parents was born in a foreign country.

School controls include the total number of enrolled students, the percentage of girls enrolled, the student-teacher ratio, as well as dummy variables indicating urban schools and privately run schools. Moreover, although PISA does not provide objective measures of the school financial situation, school resources are proxied by the school principals' subjective assessments of whether a lack of instructional material or a lack of computers hindered instruction at their school.

# 6 Results

## 6.1 DiD results

The results obtained estimating different specifications of equations (8) and (9) on different samples are reported in Tables 2 to 4. Within each table, the results are organized in panels, where the dependent variables are standardized test scores in reading, mathematics, and science, respectively.[8] Standard errors are clustered on the state level to account

[8]Estimation is performed according to the procedure recommended in OECD (2012). For each domain, OLS regressions are run separately on each of the five plausible values, and the results aggregated to obtain the final estimated coefficients and their respective standard errors. Plausible values are standardized to have mean zero and variance one in the population of ninth graders from each PISA cycle.

for serial error correlation within states over time.[9] In all instances, final sample weights are used to take into account the complex survey nature of PISA data (OECD, 2012).

Within each panel, the results for two types of specification are presented. The baseline specifications (columns 1-3) include only state and time fixed effects, besides the policy variables of interest. The main specifications (columns 4-6) add student and school controls to the corresponding baseline specifications.

In Table 2, for example, column 1 of panel A shows that on average the more challenging curriculum associated with the G8 reform increases reading test score by 0.073 standard deviations. In column 2 of the same panel, we use a student's high school grade repetition status as a crude measure of her initial preparation, and divide the students into two subgroups: those that repeated a grade in high school, and those that did not. Here, the more challenging G8 curriculum benefits the well-prepared students and increases their test scores by 0.098 standard deviations, but it hurts the less-prepared students, decreasing their test score by 0.256 standard deviations. Column 3 of panel A then reports the net difference between the two subgroups. Compared to the well-prepared students (i.e., those that did not repeat a high school grade), the less-prepared students suffer a loss in their test scores of 0.354 standard deviations. All these estimates are significant at the 5% level. However, given individual heterogeneity, the adjusted R-square of the baseline models is rather small and ranges between $0.028 - 0.034$, indicating significant variations at the individual level not captured by state or year fixed effects, or by the G8 reform dummy, which also varies at the state level.

Moving to the main specifications (columns 4-6) reported in panel A of Table 2, we first note that adding student and school controls does not have a qualitative impact on the estimated reform effects, which are our main focus. More specifically, the average effect of the more challenging curriculum under G8 is 0.072 (column 4), the effect on well-prepared students is an increase of 0.083, and that on less-prepared students is a decrease of 0.078 standard deviations (column 5). So, compared to the well-prepared students,

[9] Although this approach may lead to over-rejection of the null hypotheses when the number of clusters ($n$) is small (Cameron and Miller, 2015), this does not appear to be an issue in our setting (where $n = 16$ states): The *p-values* obtained from the wild cluster bootstrap procedure (Cameron et al., 2008) provide similar inferential results, available upon request.

the less-prepared students suffer a loss in their test scores of 0.161 standard deviations (column 6). Again, all these estimates are significant at least at the 10% level. What's more, adding the additional control variables improves the adjusted R-square of the main models to $0.104 - 0.105$, a sizable increase from the baseline models. From here on, we focus on the main specification models.

In panels B and C, the patterns for mathematics and science scores are similar. That is, the average effect of the more challenging G8 curriculum is an increase in standardized scores, but this average effect consists of two opposite effects. While the reform benefits the well-prepared students and increases their test scores, it has the opposite effect on less-prepared students, decreasing their standardized scores. Consistent with of our model predictions, the performance gap between the two subgroups of students becomes significantly larger after the reform.

As a sensitivity test, we repeat the same DiD analysis using different sample periods. Recall the main subject tested was reading in 2000 (first PISA cycle), mathematics in 2003, and science in 2006. In each case, the test for the main subject was significantly redesigned in the associated years. As a consequence, while reading tests are comparable across all PISA cycles, test comparability across cycles is ensured for math and science only since 2003 and 2006, respectively. Accordingly, we assess the robustness of our results to the exclusion of PISA 2000 from the math sample, and of PISA 2000 and 2003 from the science sample. The results are reported in Table 3. It is reassuring to see that the estimation results remain both qualitatively and quantitatively similar to those obtained using the full sample. This suggests that potential changes in the test design are not the main driver behind the seen reform effects. Hereinafter we use the full sample for estimation to achieve better efficiency.

It is also worth pointing out that, in the DiD analysis, we only rely on high school grade retention to divide students into two subgroups and separately estimate a reform effect for each subgroup. High school grade retention itself is not used as a control variable. We make this decision for the following reasons. First, as discussed before, high school grade retention is likely to lead to an upward bias in student test score, because retained

students have one more year of schooling compared to their non-retained counterparts. Our approach better insulates the potential upward bias for the subgroup of retained students, with little risk of it spilling over to the subgroup of non-retained students. Second, to the extent that test scores for a given student at different grades are correlated, adding high school grade retention on the right-hand side may lead to reverse causation. That is, instead of grade retention having an impact on the current test score, it is a student's past performance (which is correlated with his current test score) having an impact on grade retention. Again, our approach minimize the reverse causation problem by not using high school grade retention as a right-hand side control.

Nevertheless, as a further robustness check, we repeat the same DiD analysis including high school grade retention as an explicit control variable. The results are reported in Table 4. As expected, the coefficient on high school grade retention is negative, even though the interpretation can be ambiguous given the reverse causation concern. On the other hand, we still find a similar average effect of a significant increase in test scores ranging from 0.060 to 0.079 standard deviations, depending on the subjects (column 4). The impact on well-prepared students is a more pronounced increase ranging from 0.068 to 0.090 standard deviations, while that on less-prepared students is an insignificant decrease (column 5). However, compared to the well-prepared students, the less-prepared students still suffer a loss in their test scores ranging from 0.103 to 0.153 standard deviations, significant at the 5% level.

## 6.2 Quantile regression results

Next, we turn to the quantile analysis to estimate the potentially heterogeneous effects of the more challenging curriculum under G8 at different quantiles of the student standardized test score distribution.

Table 5 reports the QDiD results at all deciles of the distribution using the main specification, namely with student and school controls. Panel A reports the reform effects on reading test scores. Recall from Table 2 (column 4) that the average reform effect is 0.072, but this effect is not uniform across students. Instead, there are important distri-

butional differences. Conditioning on the observed variables, we find that the G8 reform is insignificant at the first two deciles, becomes significant at the 10% level at the third decile, and is significant at the 5% level from the fourth decile upward. Furthermore, the magnitude of the reform effect, when significant, also increases from 0.055 (3rd decile) to 0.101 (9th decile). Since these quantile regressions can only be estimated separately instead of jointly, we cannot obtain the covariance matrix across the quantiles to formally test whether these estimates are significantly different from one another. However, the pattern does appear consistent with our theoretical prediction that better-prepared students benefit more from a more challenging curriculum, in that the reform effect is increasing as we move up the deciles of the distribution. Again, since less-prepared students are more likely to experience grade repetition and at the same time more likely to locate on the lowest deciles of the distribution, the true reform effect at the lowest deciles may be confounded by the upward bias associated with grade repetition.

Similar patterns also show up in mathematics (panel B) and science (panel C) test scores. In mathematics, the reform effect is statistically insignificant at the first three deciles, and becomes significant from the fourth decile upward. When significant, the reform effect increases from 0.052 (4th decile) to 0.082 (9th decile). Similarly, in science, the reform effect is insignificant at the first decile, and becomes significant from the second decile upward. In term of magnitude, when significant, the reform effect increases from 0.064 (2nd decile) to 0.103 (9th decile). Overall, despite some minor local fluctuations, the overall pattern appears increasing as we move from left to right over the deciles.

Table 6 reports the RIF-DiD results at selected quantiles, relaxing the common distribution assumption. Again panel A uses reading test scores as the outcome variable. Interestingly, the RIF-DiD estimates exhibit a pattern qualitatively similar to that in the QDiD estimates. It is insignificant at the first decile and becomes significant from the second decile upward. When significant, the reform effect increases from 0.071 (2nd decile) to 0.101 (9th decile).

However, when mathematics test scores are considered (panel B), the pattern changes. The RIF-DiD estimates are statistically insignificant at the lowest two and the highest two

deciles, but significant in the middle of the distribution. Furthermore, the reform effect appears increasing from the left tail to the median, and then decreasing from the median to the right tail. As discussed before, RIF-DiD gives us the unconditional treatment effect, and it captures both the within-group difference and between-group difference, where groups are defined by their heterogeneity in the observed control variables. Thus, similar to Firpo et al. (2009), what we find is that while the conditional treatment effect (given by QDiD estimates) in mathematics is broadly monotonic as we move up the deciles, the unconditional treatment effect (given by RIF-DiD estimates) exhibit a non-monotonic relationship. In our case, the more challenging curriculum associated with the G8 reform widens the performance gap across students depending on their preparation levels, holding everything else constant. At the same time, it also reduces the performance gap for students with different observed heterogeneity, for example, allowing well-prepared students with disadvantaged family background in treated states to catch up with well-prepared students with advantaged family background in control states.

In panel C, the RIF-DiD result using science test scores is insignificant at the first two deciles, and becomes significant from the third decile onward. When significant, the reform effect is essentially flat and fluctuates between 0.093 (3rd decile) to 0.097 (9th decile). This seems to suggest that, while within-group difference due to the curriculum change under the G8 reform leads to similar increases in the performance gap across the deciles in all subjects, between-group difference plays a more important role in mathematics, and to a lesser extent in science, while its impact in reading is rather minimal.

Last, for easy visual comparison, we also graph the QDiD and the RIF-DiD results at percentiles of the distribution using reading (Figure 2), mathematics (Figure 3), and science (Figure 4) test scores. The solid line represents the point estimates at all percentiles, and the dashed lines represent the 95% confidence interval associated with the estimates. Since the standard error cannot be as precisely estimated at the tails of the distribution as that in the middle, it is not surprising that the confidence interval gets wider at the tails, leading to statistical insignificance of the results. Nonetheless, it can be seen that the overall pattern in the QDiD results is increasing, while that in the RIF-DiD

results is relatively flat across the three subjects.

# 7 Conclusion

The horizontal feature of the education curriculum is an important component of the education technology, yet so far it has been largely overlooked in the literature. This paper is our first step toward understanding the role of education curriculum in influencing student academic outcomes. We propose a theory of education curriculum and empirically test its predictions, using the quasi-natural experiment of the G8 reform for identification. The evidence we find, namely heterogeneous reform effects depending on student initial preparation, is broadly consistent with our theory. While the average effect of the G8 reform is an increase in student test scores, such a benefit is much more pronounced for well-prepared students. In contrast, less-prepared students (i.e., those at the left tail of the distribution) do not seem to benefit from the G8 reform at all, and they may even suffer lower test scores as a result.

In future research, we envision to extend the current paper in a couple of directions. First, our current analysis assumes that the education curriculum and other measures of school quality such as class size and teacher quality are additively separable, while the interaction between the horizontal and the vertical features of the education technology can play an important role in determining student outcomes. One possible extension is to explicitly model such interaction between the horizontal and vertical treats of the education technology. With such a model, the distributional effect of a curriculum change can depend not only on student preparation, but also on the vertical measures of school quality. Second, our analysis assumes a constant level of student effort, which again can change depending on a student's objective. For example, when a well-prepared student faces a more challenging curriculum, he may increase her study effort because the effectiveness of her learning has improved with the better-matched curriculum. Alternatively, he may also decrease her study effort if all he cares about is meeting a target test score for high school graduation or college admission, which requires less effort now that the

effectiveness of her learning has improved. Such endogenous adjustment of student effort may strengthen or weaken the distributional effect of a curriculum change, depending on whether students view their effort and the education curriculum as complements or substitutes. Extensions in these directions will help us better understand education curriculum as a critical component of the education technology, and its impact on student achievement.

# References


Aaronson, D., L. Barrow, and W. Sander (2007) 'Teachers and student achievement in the chicago public high schools.' *Journal of Labor Economics* 25(1), 95–135

Andrietti, Vincenzo (2015) 'The causal effects of increased learning intensity on student achievement: Evidence from a natural experiment.' UC3M WP Economic Series 15-06, Universidad Carlos III de Madrid, June

_ (2016) 'The causal effects of an intensified curriculum on cognitive skills: Evidence from a natural experiment.' UC3M WP Economic Series 16-06, Universidad Carlos III de Madrid, April

Angrist, Joshua D., and Kevin Lang (2004) 'Does school integration generate peer effects? evidence from boston's metco program.' *The American Economic Review* 94, 1613–1634

Angrist, Joshua D., and Victor Lavy (1999) 'Using maimonides' rule to estimate the effect of class size on children's academic achievement.' *The Quarterly Journal of Economics* 114(2), 533–575

_ (2001) 'Does teacher training affect pupil learning? evidence from matched comparisons in jerusalem public schools.' *Journal of Labor Economics* 19(2), 343–369

Arcidiacono, P., Esteban M. Aucejoz, and V. Joseph Hotz (2016) 'University differences in the graduation of minorities in stem fields: Evidence from california.' *The American Economic Review* 3(106), 525–562

Arcidiacono, Peter, and Sean Nicholson (2005) 'Peer effects in medical school.' *Journal of Public Economics* 89(2-3), 327–350

Arcidiacono, Peter, Esteban M. Aucejoz, Hanming Fang, and Ken Spenner (2011) 'Does affirmative action lead to mismatch? a new test and evidence.' *Quantitative Economics* 2(3), 303–333

Baumert, J., C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann, and M. Weiss (2009) *Programme for International Sudent Assessment 2000 (PISA 2000). Version: 1* (IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2000_v1)

Ben-Porath, Yoram (1967) 'The production of human capital and the life cycle of earnings.' *Journal of Political Economy* 75(1), 352–365

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) 'How much should we trust differences-in-differences estimates?' *The Quarterly Journal of Economics* 119(1), 249–275

Betts, Julian R. (2008) 'The impact of educational standards on the level and distribution of earnings.' *The American Economic Review* 88(1), 266–275

Blankenau, William (2005) 'Public schooling, college subsidies and growth.' *Journal of Economic Dynamics and Control* 29(3), 487–507

Blankenau, William, Steven P. Cassou, and Beth F. Ingram (2007) 'Allocating government education expenditures across k-12 and college education.' *Economic Theory* 31(1), 85–112

Borah, Bijan J., and Anirban Basu (2013) 'Highlighting differences between conditional and unconditional quantile regression approaches through an application to assess medical adherence.' *Health Economics* 22(9), 1052–1070

Cameron, Colin A., and Douglas L. Miller (2015) 'A practitioner's guide to cluster-robust inference.' *Journal of Human Resources* 50(2), 317–372

Cameron, Colin A., Jonah G. Gelbach, and Douglas L. Miller (2008) 'Bootstrap-based improvements for inference with clustered errors.' *The Review of Economics and Statistics* 90, 414–427

Card, David, and Alan B. Krueger (1992) 'Does school quality matter? returns to education and the characteristics of public schools in the united states.' *Journal of Political Economy* 107(1), 151–200

Carrell, Scott E., and James E. West (2010) 'Does professor quality matter? evidence from random assignment of students to professors.' *Journal of Political Economy* 118(3), 409–432

Carrell, Scott E., Frederick V. Malmstrom, and James E. West (2008) 'Peer effects in academic cheating.' *Journal of Human Resources* 43(1), 173–207

Carrell, Scott E., Richard L. Fullerton, and James E. West (2009) 'Does your cohort matter? measuring peer effects in college achievement.' *Journal of Labor Economics* 27(3), 439–464

Chetty, Ray, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Withmore Schanzenbach, and Danny Yagan (2011) 'How does your kindergarten classroom affect your earnings? evidence from project star.' *The Quarterly Journal of Economics* 126(4), 1593–1660

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor (2006) 'Teacher-student matching and the assessment of teacher effectiveness.' *Journal of Human Resources* 41(4), 778–820

Costrell, Robert M. (1994) 'A simple model of educational standards.' *The American Economic Review* 84(4), 956–971

_ (1997) 'Can centralized educational standards raise welfare?' *Journal of Public Economics* 65(3), 271–293

Cunha, Flavio, and James J. Heckman (2007) 'The technology of skill formation.' *The American Economic Review* 97(2), 31–47

Currie, Janet, and Thomas Dee (1995) 'Does head start make a difference?' *The American Economic Review* 85(3), 341–364

_ (2000) 'School quality and the longer-term effects of head start.' *Journal of Human Resources* 35(4), 755–774

Ding, Weili, and Steven F. Lehrer (2010) 'Estimating treatment effects from contaminated multiperiod education experiments: the dynamic impacts of class size reductions.' *The Review of Economics and Statistics* 92(1), 31–42

Driskill, Robert A., and Andrew W. Horowitz (2002) 'Investment in hierarchical human capital.' *Review of Development Economics* 6(1), 48–58

Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011) 'Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya.' *The American Economic Review* 101(5), 1739–1774

Eisenkopf, Gerald, and Ansgar Wohlschlegel (2012) 'Regulation in the market for education and optimal choice of curriculum.' *Journal of Urban Economics* 71(1), 53–65

Epple, Dennis, and Richard E. Romano (1998) 'Competition between private and public schools, vouchers, and peer-group effects.' *The American Economic Review* 88(1), 33–62

_ (2008) 'Educational vouchers and cream skimming.' *International Economic Review* 49(4), 1395–1435

Epple, Dennis, David Figlio, and Richard E. Romano (2004) 'Competition between private and public schools: testing stratification and price predictions.' *Journal of Public Economics* 88(7-8), 1215–1245

Epple, Dennis, Elizabeth Newlon, and Richard E. Romano (2002) 'Ability tracking, school competition, and the distribution of educational benefits.' *Journal of Public Economics* 83(1), 1–48

Epple, Dennis, Richard E. Romano, and Holger Sieg (2006) 'Admission, tuition, and financial aid policies in the market for higher education.' *Econometrica* 74(4), 885–928

Evans, William N., Wallace E. Oates, and Robert M. Schwab (1992) 'Measuring peer group effects: A study of teenage behavior.' *Journal of Political Economy* 100(5), 966–991

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux (2009) 'Unconditional quantile regressions.' *Econometrica* 77(3), 953–973

Gilpin, Gregory, and Michael Kaganovich (2012) 'The quantity and quality of teachers: Dynamics of the trade-off.' *Journal of Public Economics* 96(3-4), 417–429

Hanushek, Eric A. (1997) 'Assessing the effects of school resources on student performance: An update.' *Educational Evaluation and Policy Analysis* 19(2), 141–64

_ (2006) 'School resources.' In *Handbookof the Economics of Education (Vol. 2),* ed. E. A. Hanushek and F. Welch (Amsterdam: Elsevier)

Havnes, Tarjei, and Magne Mogstad (2015) 'Is universal child care leveling the playing field?' *Journal of Public Economics* 127, 100–114

Hoxby, Caroline (2000) 'The effects of class size on student achievement: New evidence from population variation.' *The Quarterly Journal of Economics* 115(4), 1239–1285

Jacob, Brian A., and Lars Lefgren (2004) 'Remedial education and student achievement: A regression-discontinuity analysis.' *The Review of Economics and Statistics* 86(1), 226–244

Kaganovich, Michael, and Xuejuan Su (2015) 'College expansion and curriculum choice.' CESifo Working Paper 5299, CESifo

Klieme, E., C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, and P. Stanat (2013) *Programme for International Sudent Assessment 2009 (PISA 2009). Version: 1* (IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2009_v1)

Krueger, Alan B. (2003) 'Economic considerations and class size.' *Economic Journal* 113, F34–F63

Kühn, Svenja M., Isabell van Ackeren, Gabriele Bellenberg, Christian Reintjes, and Grit im Brahm (2013) 'Wie viele schuljahre bis zum abitur? eine multiperspektivische standortbestimmung im kontext der aktuellen schulzeitdebatte.' *Zeitschrift für Erziehungswissenschaft* 16, 115–136

Light, Audrey, and Wayne Strayer (2000) 'Determinants of college completion: School quality or student ability?' *Journal of Human Resources* 35(2), 299–332

Lucas, Robert (1988) 'On the mechanics of economic development.' *Journal of Monetary Economics* 22(1), 3–42

Lyle, David S. (2007) 'Estimating and interpreting peer and role model effects from randomly assigned social groups at west point.' *The Review of Economics and Statistics* 89(2), 289–299

Maclean, Johanna Catherine, Douglas A. Webber, and Joachim Marti (2014) 'An application of unconditional quantile regression to cigarette taxes.' *Journal of Policy Analysis and Management* 33(1), 188–210

Mueller, Steffen (2013) 'Teacher experience and the class size effect: Experimental evidence.' *Journal of Public Economics* 98, 44–52

OECD (2012) *PISA 2009 technical report* (OECD Publishing)

Prenzel, M., C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, and R. Pekrun (2010) *Programme for International Sudent Assessment 2006 (PISA 2006). Version: 1* (IQB - Institut zur Qualitätentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2006_v1)

Prenzel, M., C. Sälzer, E. Klieme, O. Köller, J. Mang, J.-H. Heine, A. Schiepe-Tiska, and K. Müller (2015) *Programme for International Sudent Assessment 2012 (PISA 2012). Version: 1* (IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2012_v1)

Prenzel, M., J. Baumert, W. Blum, R. Lehmann, D. Leuner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, and U. Schiefele (2007) *Programme for International Sudent Assessment 2003 (PISA 2003). Version: 1* (IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2003_v1)

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain (2005) 'Teachers, schools, and academic achievement.' *Econometrica* 73(2), 417–458

Rothschild, Michael, and Lawrence J. White (1995) 'The analytics of the pricing of higher education and other services in which the customers are inputs.' *Journal of Political Economy* 103(3), 573–586

Rothstein, Jesse (2010) 'Teacher quality in educational production: tracking, decay, and student achievement.' *The Quarterly Journal of Economics* 125(1), 175–214

Sacerdote, Bruce (2001) 'Peer effects with random assignment: Results for dartmouth roommates.' *The Quarterly Journal of Economics* 116(2), 681–704

Su, Xuejuan (2004) 'The allocation of public funds in a hierarchical educational system.' *Journal of Economic Dynamics and Control* 28(12), 2485–2510

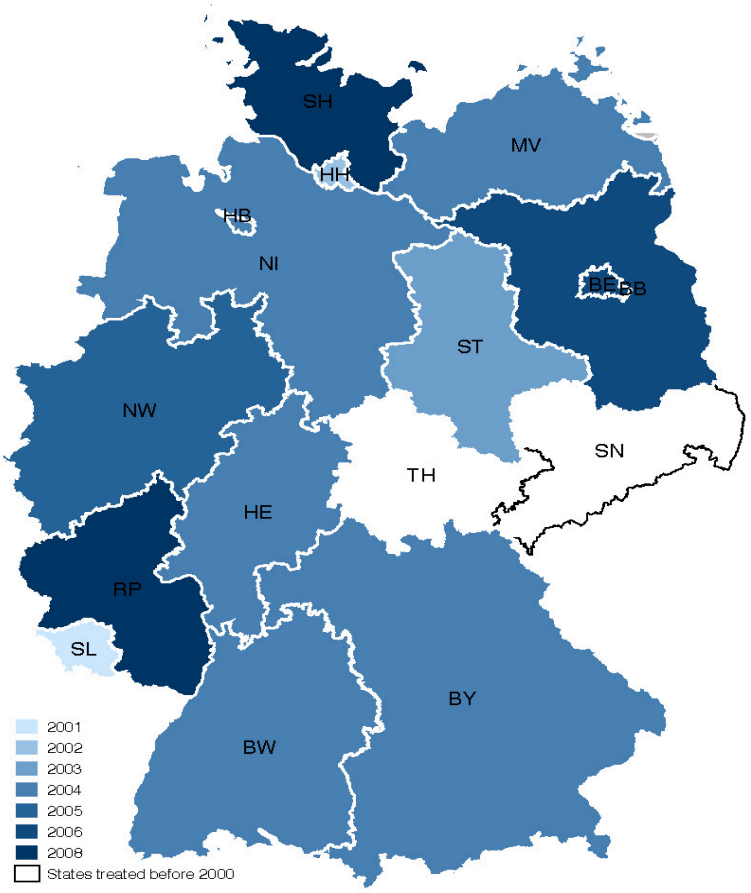
_ (2006) 'Endogenous determination of public budget allocation across education stages.' *Journal of Development Economics* 81(2), 438–456

Wiater, W. (1996) 'Zwölf jahre bis zum abitur? positionen im streit um die verk§rzung der gymnasialen schulzeit.' In *Schulreform in der Mitte der 90er Jahre: Strukturwandel und Debatten um die Entwicklung des Schulsystems in Ost- und Westdeutschland,* ed. W. Melzer and K.-J. Tillmann (Opladen: Leske + Budrich) pp. 121–139

Winston, Gordon C. (1999) 'Subsidies, hierarchy and peers: The awkward economics of higher education.' *Journal of Economic Perspectives* 13(1), 13–36

Zimmerman, David J. (2003) 'Peer effects in academic outcomes: Evidence from a natural experiment.' *The Review of Economics and Statistics* 85(1), 9–23

**Fig. 1.** Timing of the G8 reform implementation




**Legenda**
**BW**: Baden-Württemberg
**BY**: Bavaria
**BE**: Berlin
**BB**: Brandenburg
**HB**: Bremen
**HH**: Hamburg
**HE**: Hessen
**MV**: Mecklenburg-Vorpommern
**NI**: Lower Saxony
**NW**: North Rehin-Westfalia
**RP**: Rheinland-Palatinate
**SL**: Saarland
**SN**: Saxony
**ST**: Saxony-Anhalt
**SH**: Schleswig-Holstein
**TH**: Thuringia

**Table 1.** Summary statistics

| Variable | Mean | SD |
|---|---|---|
| **PISA scores** | | |
| Reading | 573.19 | 55.17 |
| Mathematics | 579.66 | 58.14 |
| Science | 588.08 | 60.89 |
| **Student controls:** | | |
| Female | 0.54 | 0.50 |
| Age (in months) | 185.19 | 5.47 |
| High school grade repeated | 0.08 | 0.26 |
| Parents' ISCED 3-4 | 0.29 | 0.46 |
| Parents' ISCED 5-6 | 0.63 | 0.48 |
| Parents' ISEI | 58.63 | 16.54 |
| Books in house: >100 | 0.60 | 0.49 |
| Only child | 0.31 | 0.46 |
| Kid born in foreign country | 0.04 | 0.20 |
| Parents born in foreign country | 0.13 | 0.34 |
| No German spoken at home | 0.04 | 0.20 |
| **School controls:** | | |
| School enrollment | 799.08 | 350.33 |
| % of girls enrolled | 49.57 | 14.89 |
| Student-teacher ratio | 14.66 | 5.88 |
| Lack of computers | 0.34 | 0.47 |
| Lack of textbooks | 0.23 | 0.42 |
| Urban school | 0.26 | 0.44 |
| Private school | 0.07 | 0.26 |
| **Policy variables:** | | |
| G8 reform | 0.41 | 0.49 |
| Observations | 31,383 | |

Notes: The sample includes academic-track ninth-graders from PISA 2000-2012 pooled data with a valid assessment in reading and non-missing values on grade retention.

**Table 2.** DiD regressions: main samples

| | Baseline | | | Main | | |
|---|---|---|---|---|---|---|
| | **(1)** | **(2)** | **(3)** | **(4)** | **(5)** | **(6)** |
| **Panel A: Reading** | | | | | | |
| G8 | 0.073** | | 0.098** | 0.072** | | 0.083** |
| | (0.021) | | (0.018) | (0.022) | | (0.020) |
| G8 × high school grade not repeated | | 0.098** | | | 0.083** | |
| | | (0.018) | | | (0.020) | |
| G8 × high school grade repeated | | -0.256** | -0.354** | | -0.078* | -0.161** |
| | | (0.037) | (0.033) | | (0.041) | (0.035) |
| Adjusted $R^2$ | 0.028 | 0.034 | 0.034 | 0.104 | 0.105 | 0.105 |
| Observations | | | 31,383 | | | |
| **Panel B: Math** | | | | | | |
| G8 | 0.069* | | 0.091** | 0.061* | | 0.072** |
| | (0.040) | | (0.038) | (0.033) | | (0.032) |
| G8 × high school grade not repeated | | 0.091** | | | 0.072** | |
| | | (0.038) | | | (0.032) | |
| G8 × high school grade repeated | | -0.217** | -0.307** | | -0.079* | -0.150** |
| | | (0.045) | (0.029) | | (0.041) | (0.030) |
| Adjusted $R^2$ | 0.031 | 0.036 | 0.036 | 0.138 | 0.139 | 0.139 |
| Observations | | | 27,381 | | | |
| **Panel C: Science** | | | | | | |
| G8 | 0.085** | | 0.109** | 0.080** | | 0.093** |
| | (0.022) | | (0.022) | (0.019) | | (0.019) |
| G8 × high school grade not repeated | | 0.109** | | | 0.093** | |
| | | (0.022) | | | (0.019) | |
| G8 × high school grade repeated | | -0.225** | -0.334** | | -0.098** | -0.190** |
| | | (0.039) | (0.038) | | (0.036) | (0.032) |
| Adjusted $R^2$ | 0.028 | 0.033 | 0.033 | 0.114 | 0.116 | 0.116 |
| Observations | | | 27,661 | | | |
| State fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Cohort fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Student controls | | | | ✓ | ✓ | ✓ |
| School controls | | | | ✓ | ✓ | ✓ |

Notes: Specifications (1)-(3) are baseline specifications. Specifications (4)-(6) are main specifications, including student and school controls. The main specifications do not include high school grade retention among the controls. Final student weights are used in all regressions. Standard errors clustered on state are reported in parentheses. ∗∗ and ∗ indicate significance at 5 and 10 percent levels, respectively. The samples in panel A, B, and C include academic-track ninth-graders from the pooled PISA 2000-2012 dataset with a valid assessment in either reading, math, or science, respectively, and with non-missing values on grade retention.

**Table 3.** DiD regressions: truncated samples

| | Baseline | | | Main | | |
|---|---|---|---|---|---|---|
| | **(1)** | **(2)** | **(3)** | **(4)** | **(5)** | **(6)** |
| **Panel B: Math** | | | | | | |
| G8 | 0.081** | | 0.103** | 0.077** | | 0.087** |
| | (0.040) | | (0.039) | (0.033) | | (0.033) |
| G8 × high school grade not repeated | | 0.103** | | | 0.087** | |
| | | (0.039) | | | (0.033) | |
| G8 × high school grade repeated | | -0.204** | -0.306** | | -0.059 | -0.145** |
| | | (0.044) | (0.029) | | (0.041) | (0.031) |
| Adjusted $R^2$ | 0.033 | 0.038 | 0.038 | 0.140 | 0.141 | 0.141 |
| Observations | | | 23,036 | | | |
| **Panel C: Science** | | | | | | |
| G8 | 0.095** | | 0.119** | 0.103** | | 0.116** |
| | (0.025) | | (0.028) | (0.031) | | (0.033) |
| G8 × high school grade not repeated | | 0.119** | | | 0.116** | |
| | | (0.028) | | | (0.033) | |
| G8 × high school grade repeated | | -0.215** | -0.334** | | -0.064 | -0.179** |
| | | (0.040) | (0.040) | | (0.045) | (0.041) |
| Adjusted $R^2$ | 0.028 | 0.036 | 0.036 | 0.130 | 0.131 | 0.131 |
| Observations | | | 15,736 | | | |
| State fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Cohort fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Student controls | | | | ✓ | ✓ | ✓ |
| School controls | | | | ✓ | ✓ | ✓ |

Notes: Specifications (1)-(3) are baseline specifications. Specifications (4)-(6) are main specifications, including student and school controls. The main specifications do not include high school grade retention among the controls. Final student weights are used in all regressions. Standard errors clustered on state are reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. The sample in panel B (C) includes academic-track ninth-graders from PISA 2003-2012 (2006-2012) with a valid assessment in math (science) and non-missing values on grade retention.

**Table 4.** DiD regressions: main samples

| | Baseline | | | Main | | |
|---|---|---|---|---|---|---|
| | **(1)** | **(2)** | **(3)** | **(4)** | **(5)** | **(6)** |
| **Panel A: Reading** | | | | | | |
| G8 | 0.072** | | 0.084** | 0.071** | | 0.081** |
| | (0.020) | | (0.020) | (0.022) | | (0.020) |
| High school grade repeated | -0.245** | -0.183** | -0.183** | -0.087** | -0.037 | -0.037 |
| | (0.030) | (0.064) | (0.064) | (0.040) | (0.059) | (0.059) |
| G8 × high school grade not repeated | | 0.084** | | | 0.081** | |
| | | (0.020) | | | (0.020) | |
| G8 × high school grade repeated | | -0.088 | -0.172** | | -0.052 | -0.134** |
| | | (0.078) | (0.083) | | (0.065) | (0.059) |
| Adjusted $R^2$ | 0.036 | 0.037 | 0.037 | 0.104 | 0.105 | 0.105 |
| Observations | | | 31,383 | | | |
| **Panel B: Math** | | | | | | |
| G8 | 0.067* | | 0.075** | 0.060* | | 0.068** |
| | (0.038) | | (0.037) | (0.032) | | (0.031) |
| High school grade repeated | -0.236** | -0.195** | -0.195** | -0.103** | -0.065* | -0.065* |
| | (0.019) | (0.035) | (0.035) | (0.032) | (0.040) | (0.040) |
| G8 × high school grade not repeated | | 0.075** | | | 0.068** | |
| | | (0.037) | | | (0.031) | |
| G8 × high school grade repeated | | -0.038 | -0.113** | | -0.035 | -0.103** |
| | | (0.064) | (0.050) | | (0.051) | (0.036) |
| Adjusted $R^2$ | 0.039 | 0.039 | 0.039 | 0.139 | 0.139 | 0.139 |
| Observations | | | 27,381 | | | |
| **Panel C: Science** | | | | | | |
| G8 | 0.083** | | 0.095** | 0.079** | | 0.090** |
| | (0.021) | | (0.022) | (0.019) | | (0.019) |
| High school grade repeated | -0.230** | -0.173** | -0.173** | -0.108** | -0.052 | -0.052 |
| | (0.021) | (0.030) | (0.030) | (0.030) | (0.036) | (0.036) |
| G8 × high school grade not repeated | | 0.095** | | | 0.090** | |
| | | (0.022) | | | (0.019) | |
| G8 × high school grade repeated | | -0.067 | -0.162** | | -0.063 | -0.153** |
| | | (0.044) | (0.046) | | (0.038) | (0.036) |
| Adjusted $R^2$ | 0.035 | 0.036 | 0.036 | 0.115 | 0.116 | 0.116 |
| Observations | | | 27,661 | | | |
| State fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Cohort fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Student controls | | | | ✓ | ✓ | ✓ |
| School controls | | | | ✓ | ✓ | ✓ |

Notes: Specifications (1)-(3) are baseline specifications. Specifications (4)-(6) are main specifications, including student and school controls. The main specifications include high school grade retention among the controls. Final student weights are used in all regressions. Standard errors clustered on state are reported in parentheses. ∗∗ and ∗ indicate significance at 5 and 10 percent levels, respectively.The samples in panel A, B, and C include academic-track ninth-graders from the pooled PISA 2000-2012 dataset with a valid assessment in either reading, math, or science, respectively, and with non-missing values on grade retention.

**Table 5.** G8 policy effects: QDiD

| | Quantiles | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **0.10** | **0.20** | **0.30** | **0.40** | **0.50** | **0.60** | **0.70** | **0.80** | **0.90** |
| **Panel A: Reading** | | | | | | | | | |
| G8 | 0.026 | 0.047 | 0.055* | 0.074** | 0.086** | 0.091** | 0.100** | 0.107** | 0.101** |
| | (0.040) | (0.029) | (0.031) | (0.031) | (0.028) | (0.024) | (0.029) | (0.033) | (0.043) |
| Observations | | | | | 31,383 | | | | |
| **Panel B: Math** | | | | | | | | | |
| G8 | 0.016 | 0.019 | 0.027 | 0.052* | 0.070** | 0.092** | 0.090** | 0.092** | 0.082* |
| | (0.047) | (0.042) | (0.032) | (0.029) | (0.027) | (0.030) | (0.028) | (0.031) | (0.045) |
| Observations | | | | | 27,381 | | | | |
| **Panel C: Science** | | | | | | | | | |
| G8 | 0.049 | 0.064* | 0.055** | 0.071** | 0.084** | 0.093** | 0.104** | 0.105** | 0.103** |
| | (0.053) | (0.034) | (0.027) | (0.029) | (0.027) | (0.027) | (0.024) | (0.034) | (0.049) |
| Observations | | | | | 27,661 | | | | |

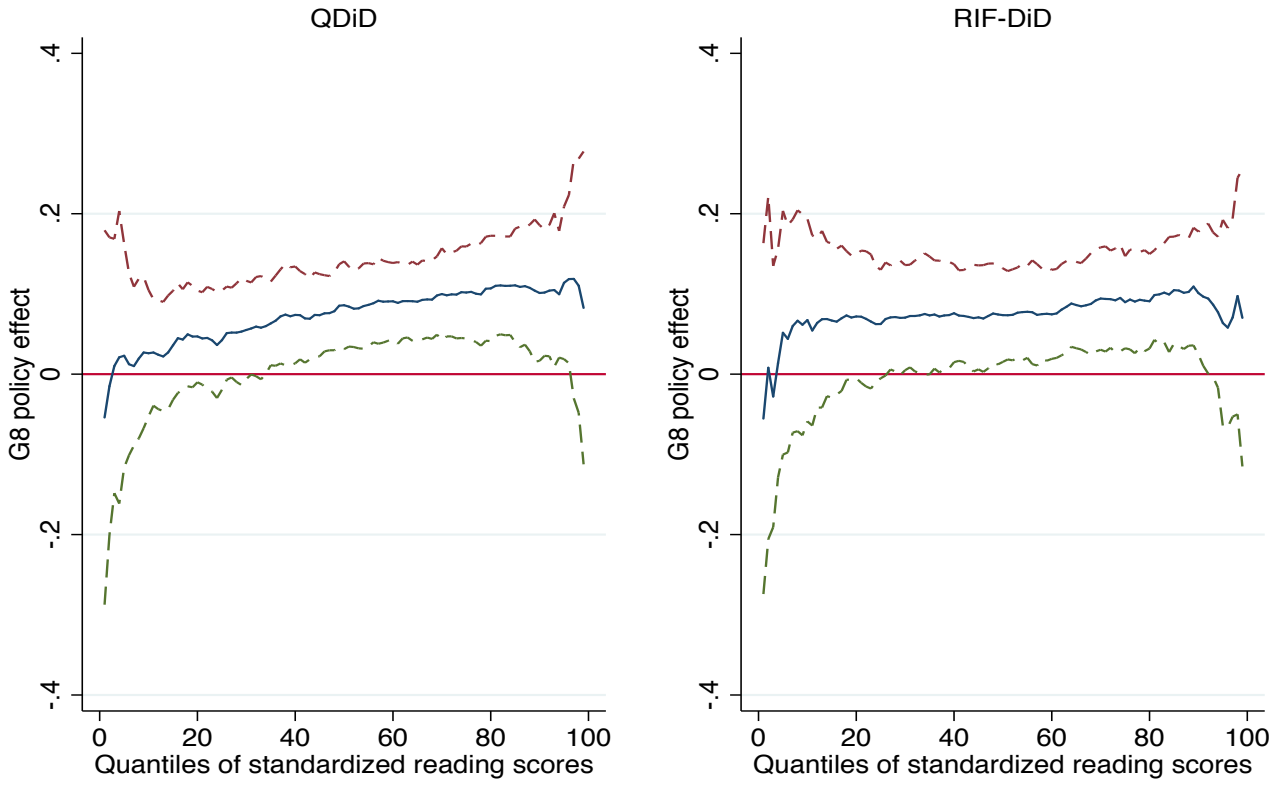
Notes: Final student weights are used in all regressions. Conventional standard errors are reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. The samples in panel A, B, and C include academic-track ninth-graders from the pooled PISA 2000-2012 dataset with a valid assessment in either reading, math, or science, respectively, and with non-missing values on grade retention.

**Table 6.** G8 policy effects: RIF-DiD

| | Quantiles | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **0.10** | **0.20** | **0.30** | **0.40** | **0.50** | **0.60** | **0.70** | **0.80** | **0.90** |
| **Panel A: Reading** | | | | | | | | | |
| G8 | 0.068 | 0.071* | 0.071** | 0.076** | 0.074** | 0.075** | 0.094** | 0.091** | 0.101** |
| | (0.066) | (0.037) | (0.033) | (0.030) | (0.029) | (0.028) | (0.034) | (0.031) | (0.039) |
| Observations | | | | | 31,383 | | | | |
| **Panel B: Math** | | | | | | | | | |
| G8 | 0.030 | 0.037 | 0.067** | 0.082** | 0.091** | 0.088** | 0.076** | 0.061 | 0.045 |
| | (0.045) | (0.039) | (0.034) | (0.030) | (0.032) | (0.034) | (0.031) | (0.043) | (0.045) |
| Observations | | | | | 27,381 | | | | |
| **Panel C: Science** | | | | | | | | | |
| G8 | 0.067 | 0.075 | 0.093** | 0.083** | 0.093** | 0.090** | 0.092** | 0.093** | 0.097* |
| | (0.066) | (0.048) | (0.034) | (0.036) | (0.032) | (0.033) | (0.032) | (0.035) | (0.051) |
| Observations | | | | | 27,661 | | | | |

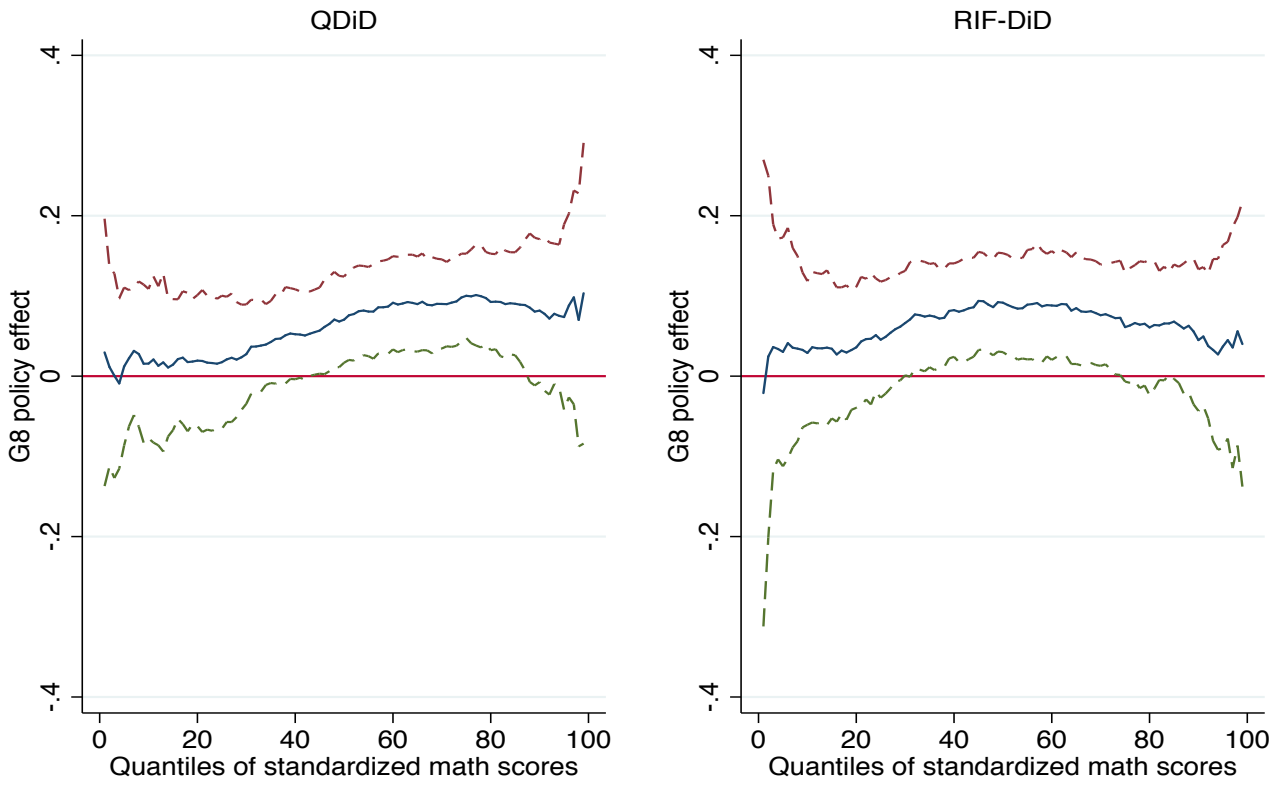
Notes: Final student weights are used in all regressions. Standard errors – reported in parentheses – are based on 200 bootstrap replications. (** and $*$ indicate significance at 5 and 10 percent levels, respectively. The samples in panel A, B, and C include academic-track ninth-graders from the pooled PISA 2000-2012 dataset with a valid assessment in either reading, math, or science, respectively, and with non-missing values on grade retention.

**Fig. 2.** G8 policy distributional effects: Reading



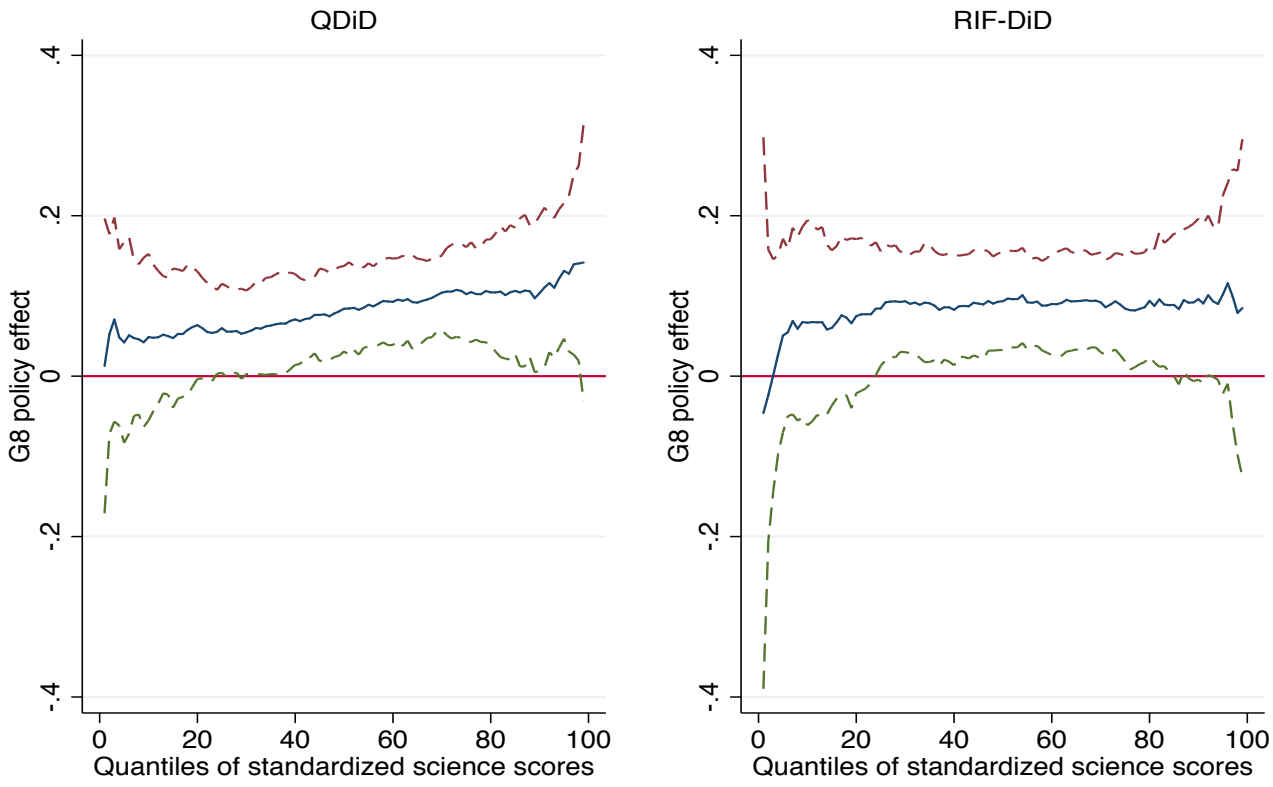

Source: Our elaborations on PISA 2000-2012 data. Rif-DiD and QDiD estimates and 95% CI

**Fig. 3.** G8 policy distributional effects: Mathematics




Source: Our elaborations on PISA 2000-2012 data. Rif-DiD and QDiD estimates and 95% CI

**Fig. 4.** G8 policy distributional effects: Science




Source: Our elaborations on PISA 2000-2012 data. Rif-DiD and QDiD estimates and 95% CI